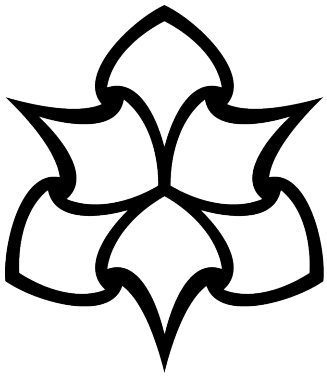

Neural Text Simplification of Clinical Letters with a Domain Specific Phrase Table

Matthew Shardlow
Department of Computing
and Mathematics
Manchester Metropolitan University
m.shardlow@mmu.ac.uk

Raheel Nawaz
Department of Operations, Technology,
Events and Hospitality Management
Manchester Metropolitan University
r.nawaz@mmu.ac.uk

Abstract

Clinical letters are infamously impenetrable for the lay patient. This work uses neural text simplification methods to automatically improve the understandability of clinical letters for patients. We take existing neural text simplification software and augment it with a new phrase table that links complex medical terminology to simpler vocabulary by mining SNOMED-CT. In an evaluation task using crowdsourcing, we show that the results of our new system are ranked easier to understand (average rank 1.93) than using the original system (2.34) without our phrase table. We also show improvement against baselines including the original text (2.79) and using the phrase table without the neural text simplification software (2.94). Our methods can easily be transferred outside of the clinical domain by using domain-appropriate resources to provide effective neural text simplification for any domain without the need for costly annotation.

1 Introduction

Text Simplification is the process of automatically improving the understandability of a text for an end user. In this paper, we use text simplification methods to improve the understandability of clinical letters. Clinical letters are written by doctors and typically contain complex medical language that is beyond the scope of the lay reader. A patient may see these if they are addressed directly, or via online electronic health records. If a patient does not understand the text that they are reading, this may cause them to be confused about their diagnosis, prognosis and clinical findings. Recently, the UK Academy of Medical Royal Colleges introduced the "Please Write to Me" Campaign, which encouraged clinicians to write directly to patients, avoid latin-phrases and acronyms, ditch redundant words and generally write in a manner that is accessible to a non-expert (Academy of Medical Royal Colleges, 2018). Inspired by this document, we took data from publicly available datasets of clinical letters (Section 3), used state of the art Neural Text Simplification software to improve the understandability of these documents (Section 4) analysed the results and identified errors (Section 5), built a parallel vocabulary of complex and simple terms (Section 6), integrated this into the simplification system and evaluated this with human judges, showing an overall improvement (Section 7).

2 Related Work

The idea of simplifying texts through machine translation has been around some time (Wubben et al., 2012; Xu et al., 2016), however with recent advances in machine translation leveraging deep learning (Wu et al., 2016), text simplification using neural networks (Wang et al., 2016; Nisioi et al., 2017; Sulem et al., 2018) has become a realistic prospect. The Neural Text Simplification (NTS) system (Nisioi et al., 2017) uses the freely available OpenNMT (Klein et al., 2017) software package[1] which provides sequence to sequence learning between a source and target language. In the simplification paradigm, the source language is difficult to understand language and the target language is an easier version of that language (in our case both English, although other languages can be simplified using the same architecture). The authors of the NTS system provide models trained on parallel data from English Wikipedia and Simple English Wikipedia which can be used to simplify source documents in English. NTS provides lexical simplifications at the level of both single lexemes and multiword expressions in addition to syntactic simplifications such as paraphrasing or removing redundant

[1] http://opennmt.net/

grammatical structures. Neural Machine Translation is not perfect and may sometimes result in errors. A recent study found that one specific area of concern was lexical cohesion (Voita et al., 2019), which would affect the readability and hence simplicity of a resulting text.

Phrase tables for simplification have also been applied in the context of paraphrasing systems where paraphrases are identified manually (Hoard et al., 1992) or learnt from corpora (Yatskar et al., 2010; Grabar et al., 2014; Hasan et al., 2016) and stored in a phrase table for later application to a text. A paraphrase consists of a complex phrase paired with one or more simplifications of that phrase. These are context specific and must be applied at the appropriate places to avoid semantic errors that lead to loss of meaning (Shardlow, 2014).

The clinical/medical domain recieves much attention for NLP (Shardlow et al., 2018; Yunus et al., 2019; Jahangir et al., 2017; Nawaz et al., 2012) and is well suited to the task of text simplification as there is a need for experts (i.e., clinicians) to communicate with non-experts (i.e., patients) in a language commonly understood by both. Previous efforts to address this issue via text simplification have focussed on (a) public health information (Kloehn et al., 2018), where significant investigations have been undertaken to understand what makes language difficult for a patient and (b) the simplification of medical texts in the Swedish language (Abrahamsson et al., 2014), which presents its own unique set of challenges for text simplification due to compound words.

3 Data Collection

To assess the impact of simplification on patient understanding, we obtained 2 datasets representing clinical texts that may be viewed by a patient. We selected data from the i2b2 shared task, as well as data from MIMIC. A brief description of each dataset, along with the preprocessing we applied is below. We selected 149 records from i2b2 and 150 from MIMIC. Corpus statistics are given in Table 1.

3.1 i2b2

The i2b2 2006 Deidentification and Smoking Challenge (Uzuner et al., 2007) consists of 889 unannotated, de-identified discharge summaries. We selected the test-set of 220 patient records and filtered these for all records containing more than 10 tokens. This gave us 149 records to work with. We concatenated all the information from each record into one file and did no further preprocessing of this data as it was already tokenised and normalised sufficiently.

	i2b2	MIMIC	Total
Records	149	150	299
Words	80,273	699,798	780,071
Avg. Words	538.7	4665.3	2,608.9

Table 1: Corpus statistics

3.2 MIMIC

In addition to i2b2, we also downloaded data from MIMIC-III v1.4 (Johnson et al., 2016) (referred to herein as MIMIC). MIMIC provides over 58,000 hospital records, with detailed clinical information regarding a patient's care. One key difference between MIMIC and i2b2 was that each of MIMIC's records contained multiple discrete statements separated by time. We separated these sub-records, and selected the 150 with the largest number of tokens. This ensured that we had selected a varied sample from across the documents that were available to us. We did not use all the data available to us due to the time constraints of (a) running the software and (b) performing the analysis on the resulting documents. We preprocessed this data using the tokenisation algorithm distributed with OpenNMT.

4 Neural Text Simplification

We used the publicly available NTS system (Nisioi et al., 2017). This package is freely available via GitHub[2]. We chose to use this rather than reimplementing our own system as it allows us to better compare our work to the current state of the art and makes it easier for others to reproduce our work. We have not included details of the specific algorithm that underlies the OpenNMT framework, as this is not the focus of our paper and is reported on in depth in the original paper, where we would direct readers. Briefly, their system uses an Encoder-Decoder LSTM layer with 500 hidden units, dropout and attention. Original words are substituted when an out of vocabulary word is detected, as this is appropriate in mono-

[2] https://github.com/senisioi/NeuralTextSimplification/

lingual machine translation. The simplification model that underpins the NTS software is trained using aligned English Wikipedia and Simple English Wikipedia data. This model is distributed as part of the software.

We ran the NTS software on each of our 299 records to generate a new simplified version of each original record. We used the standard parameters given with the NTS software as follows:

Beam Size = 5: This parameter controls the beam search that is used to select a final sentence. A beam size of 1 would indicate greedy search.

n-best = 4: This causes the 4 best translations to be output, although in practice, we only selected the best possible translation in each case.

model = NTS-w2v_epoch11_10.20.t7: Two models were provided with the NTS software, we chose the model with the highest BLEU score in the original NTS paper.

replace_unk: This parameter forces unknown words to be replaced by the original token in the sentence (as opposed to an <UNK> marker).

4.1 Readability Indices

To identify whether our system was performing some form of simplification we calculated three readability indices,[3] each of which took into account different information about the text. We have not reported formulae here as they are available in the original papers, and abundantly online.

Flesch-Kincaid: The Flesch-Kincaid reading grade calculator (Kincaid et al., 1975) takes into account the ratio of words to sentences and the ratio of syllables to words in a text. This tells us information about how long each sentence is and how many long words are used in each text. The output of Flesch-Kincaid is an approximation of the appropriate US Reading Grade for the text.

Gunning-Fox: The Gunning Fox index (Gunning, 1952) estimates the years of education required for a reader to understand a text. It takes into account the ratio of words to sentences and the proportion of words in a text which are deemed to be complex, where a complex word is considered to be any words of more than 3 syllables, discounting suffixes.

Coleman-Liau: The Coleman-Liau index (Coleman and Liau, 1975) estimates the US reading grade level of a text. It takes into account the average numbers of letters per word and sentences per word in a text.

[3] using the implementations at: https://github.com/mmautner/readability

		i2b2	MIMIC
Flesch Kincaid	Pre	8.70	6.40
	Post	6.46	4.84
	P-Value	< 0.001	< 0.001
Gunning Fox	Pre	14.53	12.69
	Post	12.35	7.36
	P-Value	< 0.001	< 0.001
Coleman Liau	Pre	10.60	10.12
	Post	9.04	5.90
	P-Value	< 0.001	< 0.001

Table 2: The results of calculating 3 readability indices on the texts before and after simplification. We show a significant reduction in the metrics in each case indicating that the texts after simplification are suitable for a lower reading grade level.

The results of each of these metrics for the i2b2 and MIMIC documents are shown in Table 2. In each case, using the NTS software improved the readability of the document. We calculated the statistical significance of this improvement with a t-test, receiving a p-value of less than 0.001 in each case. However, readability indices say nothing about the understandability of the final text and it could be the case that the resultant text was nonsensical, but still got a better score. This concern led us to perform the error analysis in the following section.

5 Error Analysis

Our previous analysis showed that the documents were easier to read according to automated indices, however the automated indices were not capable of telling us anything about the quality of the resulting text. To investigate this further, we analysed 1000 sentences (500 from i2b2 and 500 from MIMIC) that had been processed by the system and categorised each according to the following framework:

Type 1: A change has been made with no loss or alteration of the original meaning.

Type 2: No change has been made.

Type 3: A significant reduction in the information has been made, which has led to critical information being missed.

Type 4: A single lexical substitution has been made, which led to loss or alteration of the original meaning.

Type 5: An incorrect paraphrase or rewording of the sentence has been made, which led to loss or alteration of the original meaning.

Type 6: A single word from the original text is repeated multiple times in the resulting text.

We developed this framework by looking at the 1000 sentences in our corpus. Although the framework does not give any information about the readability of sentences, it does tell us about the existing pitfalls of the algorithm. We were able to categorise every sentence using these six categories. Each category represents an increased level of severity in terms of the consequences for the readability of the text. A Type 1 sentence may have a positive impact on the readability of a text.[4] A Type 2 sentence will not have any impact as no modification has been made. A Type 3 sentence may improve the readability according to the automated metric and may help the reader understand one portion of the text, however some critical information from the original text has been missed. In a clinical setting, this could lead to the patient missing some useful information about their care. Types 4, 5 and 6 represent further errors of increasing severity. In these cases, the resulting sentences did not convey the original meaning of the text and would diminish the understandability of a text if shown to a reader.

The first author of this paper went through each sentence with the categories above and assigned each sentence to an appropriate category. Where one sentence crossed multiple categories, the highest (i.e., most severe) category was chosen. However, this only occurred in a small proportion of the data and would not significantly affect our results had we recorded these separately. The results of the error analysis are shown in Table 3.

[4]note, we do not claim that all Type 1 sentences are simplifications, only that the system has made a change which is attempting to simplify the text. This may or may not result in the text being easier to understand by a reader.

Type	i2b2	MIMIC	Total
1	25	33	58
2	337	322	659
3	41	55	96
4	55	61	116
5	25	21	46
6	17	8	25

Table 3: The results of the error analysis. 500 sentences each were annotated from i2b2 and MIMIC to give 1000 annotated sentences in the 'Total' column.

The results show that the majority of the time the system does not make a change to the text ($659/1000 = 65.9\%$ of the time). We would not expect every single sentence to be simplified by the system, as some sentences may not require simplification to be understood by an end user. Other sentences may require simplification, but the system does not realise this, in which case the system may still choose not to simplify the text. Only in 5.8% of the cases is a valid simplification made. These generally consisted of helpful lexical substitutions, however there were also some examples of helpful rephrasing or paraphrasing. In addition to the 5.8% of valid simplifications, a further 9.6% of cases were instances where a significant chunk of a sentence had been removed. In these cases, the resulting sentence was still readable by an end user, however some important information was missing. These sentences do not necessarily constitute an error in the system's behaviour as the information that was omitted may not have been relevant to the patient and removing it may have helped the patient to better understand the text overall, despite missing some specific detail. The rate of Type 4 errors is 11.6%. These errors significantly obfuscated the text as an incorrect word was placed in the text, where the original word would have been more useful. 4.6% of errors were incorrect rewordings (Type 5) and a further 2.5% were cases of a word being repeated multiple times. In total this gives 18.7% of sentences that result in errors. The error rate clearly informs the use of the NTS software. It may be the case that in a clinical setting, NTS could be used as an aid to the doctor when writing a patient letter to suggest simplifications, however it is clear that it

would not be appropriate to simplify a doctor's letter and send this directly to a patient without any human intervention.

6 Phrase Table Development

The NTS system is trained on parallel Wikipedia and Simple Wikipedia documents. Whilst these may contain some medical texts, they are not specific to the clinical genre and we should not expect that direct simplification of medical language will occur. Indeed, when we examined the texts, it was clear that the majority of simplifications that were made concerned general language, rather than simplifying medical terminology. One way of overcoming this would be to create a large parallel corpus of simplified clinical letters. However this is difficult due to the licensing conditions of the source texts that we are using, where an annotator would be required to agree to the licence conditions of the dataset(s). In addition, we would require clinical experts who were capable of understanding and simplifying the texts. The clinical experts would have to produce vast amounts of simplified texts in order to provide sufficient training data for the OpenNMT system to learn from. Although this is possible, it would require significant time and financial resources.

OpenNMT provides an additional feature that allows a pre-compiled phrase table to be used when an out-of-vocabulary word is identified. This can be used in cross-lingual translation to provide idioms, loan words or unusual translations. In monolingual translation, we can use this feature to provide specific lexical replacements that will result in easier to understand text. This allows us to use a general language simplification model, with a domain-specific phrase table and effectively simplify complex vocabulary from the (clinical) domain.

We downloaded the entire SNOMED-CT clinical thesaurus (Donnelly, 2006), which contains 2,513,952 clinical terms, each associated with a concept identifier. We chose this resource over the full UMLS Metathesaurus as SNOMED-CT contains terms specific to the clinical domain and we expected this would lead to fewer false positives. Where terms share an identifier, these are considered synonymous with each other, allowing us to create groups of semantically equivalent terms. We filtered out terms that were greater than 4 tokens long or contained punctuation, As these indicated sentential terms that were not appropriate for our purposes. We identified abbreviations and automatically removed any explanations that were associated with these. We used the Google Web1T frequencies to identify which terms were the most common in general language use. Although this is not a direct measure of how easy to understand each word will be, it has been shown previously that lexical frequency correlates well with ease of understanding (Paetzold and Specia, 2016). Where there were multi-word expressions, we took the average frequency of all words in the multi-word expression, rather than taking the frequency of the N-gram. For each set of semantically equivalent terms, we took the most frequent term as the easiest to understand and added one entry to our phrase table for each of the other terms contained in the group. So, for a group of 3 terms, A, B and C, where B is the most frequent, we would add 2 pairs to our phrase table A-B, and C-B. This means that whenever A or C are seen in the original texts and they are considered to be out-of-vocabulary words, i.e., technical medical terms that were not present in the training texts, then the more frequent term B, will be substituted instead. We identified any instances where one word had more than one simplification (due to it being present in more than one synonym group). If the original word was an acronym, we removed all simplifications as an acronym may have multiple expansions and there is no way for the system to distinguish which is the correct expansion. If the original word with more than one simplification is not an acronym then we selected the most frequent simplification and discarded any others. This resulted in 110,415 pairs of words that were added to the phrase table.

In Table 4, we have shown examples of the types of simplifications that were extracted using the methodology outlined above. Clearly these are the type of simplifications that would be helpful for patients. In some cases, it may be possible that the resulting simplified term would still be difficult to understand for an end user, for example 'hyperchlorhydria' is translated to 'increased gastric acidity', where the term 'gastric' may still be difficult for an end user. A human may have simplified this to 'increased stomach acidity', which would have been easier to understand. This phrase was not in the SNOMED-CT vocabulary and so was not available for the construction of our phrase ta-

ble. Nonetheless, the type of simplifications that are produced through this methodology appear to improve the overall level of understanding of difficult medical terms.

The methodology we have outlined above is suitable for domains outside of medical terminology. The only domain-specific resource that is required is a thesaurus of terms that are likely to occur in the domain. By following the methodology we have outlined, it would be simple to create a phrase table for any domain, which could be applied to the NTS software that we have used in this work.

7 Human Evaluation

In our final section of experiments, we wanted to determine the effect that our system had on the ease of understanding of sentences from the original texts. We evaluated this through the use of human judges. In order to thoroughly evaluate our system we compared the original texts from i2b2 and MIMIC to three methods of transformation as detailed below:

Original Texts (ORIG): We used the original texts as they appeared after preprocessing. This ensured that they were equivalent to the transformed texts and that any effects would be from the system, not the preprocessing.

NTS: We ran the sentences through the NTS system using the configuration described in Section 4.

NTS + Phrase Table (NTS + PT): We ran the sentences through the NTS system. We configured OpenNMT to use the phrase table that we described in Section 6. Note that the phrase table was only used by the system when OpenNMT identified a word as being out-of-vocabulary.

Phrase Table Baseline (PTB): To demonstrate that the benefit of our system comes from using the phrase table in tandem with the NTS system, we also provided a baseline which applied the phrase table to any word that it was possible to replace in the text.

We collected the sentences for each of the methods as described above from both of our data sources and collated these so as we could compare the results. We analysed the data and removed any instances of errors that had resulted from the NTS system, according to our error analysis. The sentences that we selected correspond to Type 1, in our categorisation. Type 1 does not necessarily indicate a simplification, instead it implies that a transformation has been successfully completed, with the potential for simplification. Selecting against errors allows us to see the simplification potential of our system. We do not claim that NTS can produce error-free text, but instead we want to demonstrate that the error-free portion of the text is easier to understand when using our phrase table. We selected 50 4-tuples from each dataset (i2b2 and MIMIC) to give 100 4-tuples, where one 4-tuple contained parallel sentences from each of the methods described above. Sentences within a 4-tuple were identical, apart from the modifications that had been made by each system. No two sentences in a 4-tuple were the same. We have put an example 4-tuple in Table 5, to indicate the type of text that was contained in each.

We used crowd sourcing via the Figure Eight platform to annotate our data. As we had a relatively small dataset, we chose to ask for 10 annotations for each 4-tuple. We allowed each annotator to complete a maximum of 20 annotations to ensure that we had a wide variety of perspectives on our data. No annotator saw the same 4-tuple twice. We provided a set of test annotations, which we intended to use to filter out bad-actors, although we found that all annotators passed the test adequately. We selected for annotators with a higher than average rating on the Figure Eight platform (level 2 and above). In each annotation, we asked the annotator to rank the 4 sentences according to their ease of understanding, where the top ranked sentence (rank 1) was the easiest to understand and the bottom ranked sentence (rank 4) was the hardest to understand. We explicitly instructed annotators to rank all sentences, and to use each rank exactly once. If an annotator found 2 sentences to be of the same complexity, they were instructed to default to the order in which the sentences were displayed. We posed our task as 4 separate questions with the exact wording shown in the supplementary material, where we have reproduced the instructions we provided to our annotators. In our post analysis we identified that 20 out of the 1000 annotations that we collected (100 4-tuples, with 10 annotation per 4-tuple) did not use all 4 ranks (i.e., 2 or more sentences were at the same rank). There was no clear pattern of spamming and we

Complex Term	Simple Term
ability to be ambulant	ability to walk
carcinoma of stomach	cancer of stomach
hyperchlorhydria	increased gastric acidity
hypertension	high blood pressure
lying supine	lying on back
osteophyte	bony spur
photophobia	intolerance to light
talipes	congenital clubfoot
AACTG	aids clinical trial group
BIPLEDS	bilateral periodic epileptiform discharges
BLADES	bristol language development scale

Table 4: Term pairs that were created for our phrase table.

System	Sentence
ORIG	Patient has been suffering from photophobia and wheezing.
NTS	Patient *suffers* from photophobia and wheezing.
NTS + PT	Patient *suffers* from *sensitivity to light* and wheezing.
PTB	Patient has been suffering from *sensitivity to light* and *asthmatic breath sounds*.

Table 5: An example of the type of text produced by our system. The NTS system has performed a syntactic simplification, converting "has been suffering" to "suffers", the NTS + PT system has simplified "photophobia" to "sensitivity to light" and the baseline system (PTB) has further simplified "wheezing" to "asthmatic breath sounds".

chose to ignore these 20 sentences in our further analysis, giving us 980 rankings.

In Table 6, we have shown the raw results of our crowd sourcing annotations as well as the average rank of each system. We calculate average rank r_s of a system s as

$$r_s = \frac{\sum_{i=1}^{4} i \times f(s,i)}{\sum_{i=1}^{4} f(s,i)}$$

where i is a rank from 1 to 4 and $f(s,i)$ is a function that maps the system and rank to the number of times that system was placed at that rank (as shown in Table 6). We can see that our system using NTS and the phrase table has the highest average rank, indicating that the text it produced was the easiest to understand more often than other systems. The NTS is ranked second highest indicating that in many cases this system still produces text which is easier to understand than the original. The original texts are ranked third most frequently, ahead of the baseline system which is most often ranked in last position. The baseline system overzealously applied simplifications from our phrase table and this led to long winded explanations and words being simplified that did not require it.

System	Rank				Avg
	1	2	3	4	
NTS + PT	**430**	255	230	65	1.93
NTS	259	**294**	264	163	2.34
ORIG	120	222	**381**	257	2.79
PTB	171	209	105	**495**	2.94

Table 6: The results of our crowdsourcing annotations. We have ordered the annotations by their average rank and highlighted the most common rank for each system. The first column in the table shows the system. Columns 2 through 5 show the number of times each system was ranked at rank 1, 2, 3 or 4 and column 6 shows the average rank calculated according to the formula in Section 7

.

8 Discussion

In our work we have applied NTS software to clinical letters and adapted the software using a bespoke phrase table mined from SNOMED-CT. We have shown the types of errors that can occur when using NTS software and we have evaluated our improved algorithm against the state of the art, showing an improvement.

Our system improved over the original NTS

software when adapted to use our phrase table. The NTS software was developed by using parallel sentences from Wikipedia and Simple Wikipedia and training OpenNMT to learn simplifications from these. OpenNMT learns an internal set of vocabulary substitutions, however these will have been targeted towards general language, rather than medical specific language. By using our phrase table, we are able to give specific simplifications for medical terms. The system only accesses the phrase table when it detects a word which is out-of-vocabulary, i.e., a word that was not seen sufficiently often in the training texts to be incorporated into the model that was produced. This works well at modelling a lay reader, where the vocabulary understood by the system is analogous to the vocabulary understood by a typical (i.e., non-specialised) reader of English.

In addition to the NTS system adapted to use our phrase table, we also tested a baseline which greedily applied the phrase table at all possible points in a sentence. However, this system was ranked as least understandable more often than any other system. The text it produced was generally much longer than the original text. The benefit of our work comes from using the phrase table together with the neural text simplification software, which is capable of applying the phrase table at the correct points in the text. This can be seen in Table 5, where the NTS system has altered the language being used, but has not simplified a medical term, the NTS + PT system has simplified the medical term (photophobia), but left a term which would be generally understood (wheezing) and the baseline system has correctly simplified the difficult medical term, but has also changed the generally understood term. Our phrase table is additional to the NTS system and could be applied to other, improved neural models for text simplification as research in this field is progressed. We have shown that our phrase table adds value to the NTS system in the clinical setting.

We have demonstrated in Section 5 that the type of text produced by NTS software and by our adapted NTS software will contain errors. This is true of any translation software which relies on learning patterns from data to estimate future translations of unseen texts. In cross-lingual translation, a small error rate may be acceptable as the text is transformed from something that is initially incomprehensible to text in the reader's own language which may be intelligible to some degree. With simplification, however, even a small error rate may lead to the resulting text becoming more difficult to understand by an end user, or the meaning of a text being changed. This is particularly the case in the clinical setting, where life changing information may be communicated. It is important then to consider how to use Neural Text Simplification in a clinical setting. We would propose that the clinician should always be kept in the loop when applying this type of simplification. The system could be applied within a word editor which suggests simplifications of sentences as and when they are discovered. The clinician could then choose whether or not to accept and integrate the simplified text.

We have presented our methodology in the context of the clinical domain, however it would be trivial to apply this elsewhere. Our methodology is particularly suitable when 3 conditions are met: (a) There is text being produced by experts that is read by lay readers. (b) that text contains specialised terminology that will be unintelligible to the intended audience and (c) a comprehensive thesaurus of domain specific terms exists, which can be used to generate a domain appropriate phrase table. Given these conditions are met, our work could be applied in the legal, financial, educational or any other domain.

We have made significant use of licensed resources (i2b2, MIMIC and SNOMED-CT). These are available for research purposes from their providers, given the user has signed a licensing agreement. We are not at liberty to share these resources ourselves and this inhibits our ability to provide direct examples of the simplifications we produced in our paper. To overcome this, we have provided the following GitHub repository, which provides all of the code we used to process the data: `https://github.com/MMU-TDMLab/ClinicalNTS`. Instructions for replication are available via the GitHub.

9 Conclusion + Future Work

Our work has explored the use of neural machine translation for text simplification in the clinical domain. Doctors and patients speak a different language and we hope that our work will help them communicate. We have shown that general language simplification needs to be augmented with domain specific simplifications and that doing so

leads to an improvement in the understandability of the resulting text.

One clear avenue of future work is to apply this system in a clinical setting and to test the results with actual patients. We will look to develop software that uses NTS to identify possible simplifications for a clinician when they are writing a letter for a patient. We could also look to use parallel simplified medical text to augment the general language parallel text used in the NTS system. Additionally, we could improve the measure of lexical complexity for single and multi word expressions. Currently, we are only using frequency as an indicator of lexical complexity, however other factors such as word length, etymology, etc. may be used. Finally, we will explore adaptations of our methodology for general (non-medical) domains, e.g., simplified search interfaces (Ananiadou et al., 2013) for semantically annotated news (Thompson et al., 2017).

References

Emil Abrahamsson, Timothy Forni, Maria Skeppstedt, and Maria Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 57–65.

Academy of Medical Royal Colleges. 2018. Please, write to me. Writing outpatient clinic letters to patients.

Sophia Ananiadou, Paul Thompson, and Raheel Nawaz. 2013. Enhancing search: Events and their discourse context. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 318–334. Springer.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Natalia Grabar, Thierry Hamon, and Dany Amiot. 2014. Automatic diagnosis of understanding of medical words. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 11–20.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

Sadid A. Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. 2016. Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 42–53, Osaka, Japan. The COLING 2016 Organizing Committee.

James E Hoard, Richard Wojcik, and Katherina Holzhauser. 1992. An automated grammar and style checker for writers of simplified english. In *Computers and Writing*, pages 278–296. Springer.

M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid, and R. Nawaz. 2017. An expert system for diabetes prediction using auto tuned multi-layer perceptron. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 722–728.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Nicholas Kloehn, Gondy Leroy, David Kauchak, Yang Gu, Sonia Colina, Nicole P Yuan, and Debra Revere. 2018. Improving consumer understanding of medical text: Development and validation of a new subsimplify algorithm to automatically generate term explanations in english and spanish. *Journal of medical Internet research*, 20(8).

Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2012. Identification of manner in bio-events. In *LREC*, pages 3505–3510.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 85–91.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *LREC*, pages 1583–1590.

Matthew Shardlow, Riza Batista-Navarro, Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2018. Identification of research hypotheses and new knowledge from scientific literature. *BMC Medical Informatics and Decision Making*, 18(1):46.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 162–173.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2017. Enriching news events with meta-knowledge information. *Language Resources and Evaluation*, 51(2):409–438.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic deidentification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machinetranslation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1015–1024. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Yunus, O. Arif, H. Afzal, M. F. Amjad, H. Abbas, H. N. Bokhari, S. T. Haider, N. Zafar, and R. Nawaz. 2019. A framework to estimate the nutritional value of food in real time using deep learning techniques. *IEEE Access*, 7:2643–2652.